Decision Science Letters 12 (2023) 291–296

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

# Time series prediction of novel coronavirus COVID-19 data in west Java using Gaussian processes and least median squared linear regression

**Intan Nurma Yulita[a,c,*], Firman Ardiansyah[b], Aulia Siska[b] and Ino Suryana[c]**

[a]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran, Bandung, Indonesia*
[b]*Master of Management, Institut Teknologi dan Bisnis Ahmad Dahlan Lamongan, Indonesia*
[c]*Department of Computer Science, Universitas Padjadjaran, Bandung, Indonesia*

**C H R O N I C L E**

*Article history*:
Received: November 26, 2022
Received in revised format: December 20, 2022
Accepted: January 13, 2023
Available online: January 13, 2023

*Keywords*:
*Time series prediction*
*COVID-19*
*Gaussian Processes*
*Linear Regression*
*West Java*

**A B S T R A C T**

In 2019, the COVID-19 epidemic swept throughout the globe. The virus was first identified in Wuhan, China. By the time several months had gone by, this virus had spread to numerous locations throughout the world. Consequently, this virus has become a worldwide pandemic. Multiple efforts have been made to limit the transmission of this virus. A possible course of action is to lock down the territory. Unfortunately, this strategy wrecked the economy, worsening the terrible situation. The world health organization (WHO) would breathe a sigh of relief if there were to be no new cases. However, the government should explore employing data from the future in addition to the data it already has. Prediction of time series may be utilized for this purpose. This study indicated that the Gaussian processes method outperformed the least median squared linear regression method (LMSLR). Applying a Pearson VII-based global kernel produces MAE and RMSE values of 23.12 and 53.43, respectively.



## 1. Introduction

China reported the first instance of the severe acute respiratory syndrome (SARS) in 2002. (Hui et al., 2020). The initial SARS case in 18 years. In December of this year, the new coronavirus reappeared in China (Lai et al., 2020). Numerous nations have been afflicted with the novel coronavirus strain known as SARS-CoV-2. Since the adoption of COVID-19, a number of nations have been ineffective (Gitt et al., 2020). Similar to its predecessor in 2002, SARS-CoV-2 caused widespread anxiety and apprehension. (Guo et al., 2020). There have been several efforts and strategies created to combat the coronavirus. Nonetheless, the illness continues to spread. COVID-19 has caused a significant number of fatalities. Beyond the sphere of public health, the COVID-19 outbreak has had implications. The global economy has ceased to function since a lockdown policy was implemented (Nicola et al., 2020). It is feasible that this may cause a global disaster. In response, a number of governments have loosened lockdown regulations and let businesses reopen in light of the COVID-19 outbreak.

On March 2, 2020, Depok, West Java, Indonesia announced the first case of Ebola in the nation (Toharudin et al., 2020). Indonesia employs extensive social restrictions (PSBB) rather than a lockdown (Zuhairoh, 2020). Similar to other regions, this restriction precipitated a recession (Olivia et al., 2020). The PSBB was subsequently relaxed across Indonesia, notably in West Java. However, the world health organization (WHO) defines conditions under which an area may modify its requirements. The absence of any confirmed instances in the region is one of them. According to the available data trend, there were several days in May when there were no incidents in West Java, but this pattern altered thereafter. There has been a surge in the number of reported cases across all of Java during June, particularly in West Java.

* Corresponding author.
E-mail address: intan.nurma@unpad.ac.id (I. N. Yulita)


doi: 10.5267/dsl.2023.1.006

Computer science may assist in forecasting the future, and it is not just valuable for seeing the present number of new cases that the easing policy has revealed (Weigend, 2018). Linear regression is a prominent approach. The future may be forecasted using linear regression, a technique that evaluates the relationship between several inputs. This method is basic, and the data structure is irrelevant when employing it. Linear regression seeks to characterize the relationship between two variables using linear equations and observable data (Hayes et al., 2020). In statistical investigations, it is typical to regard one variable as an explanatory variable and the other as a dependent variable. In July, however, the number of confirmed cases of COVID-19 in West Java surged substantially. Compared to other dates, this date has a great deal of information. The possibility of an increase must be considered. If these concerns are neglected, it is feasible that the technique will give erroneous findings. When this occurs, the resulting findings are unintelligible. Therefore, the technique must be robust in this environment. Least-median-squares linear regression is one strategy in this area (Cheng et al., 2016). The algorithms for machine learning include linear regression. Machine learning aims to automate the process of constructing analytical models from acquired data. It is an area of artificial intelligence (AI) that aims to automate as many menial jobs as possible by teaching computers to spot patterns in data and form conclusions with minimal human involvement. Its application has aided in classifying botnet attacks (Wildani et al., 2019), graph clustering (Yulita et al., 2013), identifying and classifying different stages of sleep (Yulita et al., September 2017; Yulita et al., October 2017), recognizing emotions (Yulita et al., 2019), and recognizing speech (Yulita et al., 2018). Linear regression is a form of supervised approach that takes the use of data. A regression procedure is executed. Independent variables are used to model a predicted value in regression.

In actuality, there are not a great number of problems with plainly identifiable causal links among their constituent elements. It indicates that linear regression may struggle to identify it, resulting in a significant prediction error. Furthermore, the shape of the prediction model is presented using linear regression. Nonparametric given the form flexibility of real-world data, Bayesian modeling gives a solution to this challenge. The advantages of the nonparametric model are intrinsic to the model itself; in particular, the model makes no assumptions about the parametric form. The gaussian process regression model is a bayesian nonparametric model example (Richardson et al., 2017). Here, we assume that regression functions adhere to a typical normal distribution, namely the double-normal distribution. Together, the mean function and the variance function define a Gaussian process. The variance value for the model's outputs corresponding to its inputs is computed as a function of the variance function. Different hyperparameters are utilized for different purposes. This hyperparameter's precise value is unknown but may be deduced from the gathered data. A common type of function is the quadratic exponential (Gaussian) function. The kernel is utilized for parameter estimation. Using the Gaussian processes method, this study proposes the Pearson VII function as the foundation for a universal kernel.

## 2. Material and methods

This research examines two approaches for predicting time series. Gaussian Processes and Least Median Squared Linear Regression are used. Sections 2.1 and 2.2 cover these two ways.

### *2.1. Gaussian processes*

The artificial neural network is a well-known algorithm that has proven effective in a variety of fields. Gaussian processes are one of the possible methods. It employs a huge number of hidden units to provide more accurate forecasts (Sheng et al., 2017). The Gaussian process is a stochastic process whereby any finite collection of random variables, Y, can be used. It possesses a dual Gaussian distribution. It is defined as a mean and standard deviation function. Their equations are represented by Eq. (1) and Eq. (2).

$$\mu(x) = E(Y_x) \tag{1}$$

$$k(x_i . x_j) = E(Y_{x_i} - \mu(x_i))(Y_j - \mu(x_j)) \tag{2}$$

From the perspective of nonparametric Bayesian regression, time series prediction using the Gaussian process may be generated by explicitly putting the prior Gaussian distribution for the regression functions f (x) (Schulz et al., 2018). Given several observations and a variance function, for instance. In addition, the prediction value will be computed using a model of the Gaussian process. If x* is a test point and f* is a function corresponding to x*, then the common distribution of f is a zero-mean Double Gaussian. It is defined by Eq. (3).

$$\begin{bmatrix} f \\ f^* \end{bmatrix} X, 0 \sim N\left(\begin{bmatrix} K & k \\ k^T & \kappa \end{bmatrix}\right) \tag{3}$$

where K is the x-n dependent matrix of X and every I j) member of K is k (xi, xj) (Liu et al., 2020). is a vector. Eq. (4) depicts the standard distribution of the observed values y and y*.

$$\begin{bmatrix} y \\ y^* \end{bmatrix} X, 0 \sim N\left(0, \begin{bmatrix} K + \sigma^2 I & k \\ k^T & \kappa + \sigma^2 \end{bmatrix}\right) \tag{4}$$

where $\sigma^2$ is variance.

The marginal spread of y* is shown in Equation 5.

$$y^*|\, y, X, 0 \sim N(\mu(x^*), v(x^*)) \tag{5}$$

with

$$\mu(x^*) = k^T(K + \sigma^2 I)^{-1} y \tag{6}$$

$$v(x^*) = \kappa + \sigma^2 - k^T(K + \sigma^2 I)^{-1} y \tag{7}$$

The estimation for y* is m(x*), and the variance for the estimate for y* is v(x*). For m test data X* = [$x_1$*, …. $x_m$*] then the y* distribution is double Gaussian with the following parameters in Eq. (8) and Eq. (9):

$$\mu(X^*) = K^{*T}(K + \sigma^2 I)^{-1} y \tag{8}$$

$$v(X^*) = K^{**} + \sigma^2 I - K^{*T}(K + \sigma^2 I)^{-1} K \tag{9}$$

where K* is the n x m matrix of the variance between training inputs and test points, the matrix K** with size m x m is composed of the variance between the test points.

### *2.2. Least median squared linear regression (LMSLR)*

The number of new cases in West Java fluctuates. At times, there is an increase in instances that have never occurred before. Robust regression is dependable when dealing with such severe data. Least Median Squared Linear Regression (LMSLR) is one of the approaches of robust regression (Yu et al., 2017). As stated in Eq. (10), this approach determines the median square of the residual for each iteration.

$$M_j = med\ (e_i^2) \tag{10}$$

Thus, we have M1, M2,..., Ms, which represents the median of the squares remaining after each observation hi. To determine the value of M1, we search for a subset of data from many observations, as shown in Eq. (11):

$$h_i = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right] \tag{11}$$

where n is the number of data, and p is the number of parameters. In calculating the value of $h_i$, it must always be an integer (Pati, 2020). Therefore, if the value of hi is not in the form of an integer, rounding up is performed. And so on until the iteration ends at iteration i, when $h_i$=$h_i$+1. After that, look for the minimum value of M1, M2,…., Ms.

Because LMSLR is an estimator in robust regression, it is the same as other estimators in robust regression, the basic principle of LMSLR is to assign $w_{ii}$ to it so that outlier data does not affect the estimation parameter model. The weight of $w_{ii}$ is 1 if 1 jika $|e_i/\hat{\sigma}| \leq 2.5$ and 0 otherwise. $\hat{\sigma}$ is calculated based on Eq. (12).

$$\hat{\sigma} = 1.48\ [1+\tfrac{5}{(n-p)}]\sqrt{M_j min} \tag{12}$$

After the $w_{ii}$ is calculated, the matrix is shown in Eq. (13).

$$W = \begin{bmatrix} w_{11} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & 0 & w_{nn} \end{bmatrix} \tag{13}$$

By using the matrix W, the LMSLR regression parameter estimates can be calculated using Eq. (14).

$$\widehat{\theta_{LMSLR}} = (X^T W X)^{-1}(X^T W Y) \tag{14}$$

## 3. Methodology

This study employed two modeling techniques to anticipate the COVID-19 time series in West Java. Obtainable information from the website https://pikobar.jabarprov.go.id. Only daily instances from March 2, 2020 to September 5, 2020 were evaluated. It was reached on July 9, 2020. On that date, there were 965 confirmed cases. The number of new instances of corona or COVID-19 in West Java increased tenfold from the day before. West Java's rapid spread of the COVID-19 was attributed to a new cluster at the army officer candidate school (Secapa AD). Not just in West Java, but also nationally, it led to the largest increase in cases to that point.

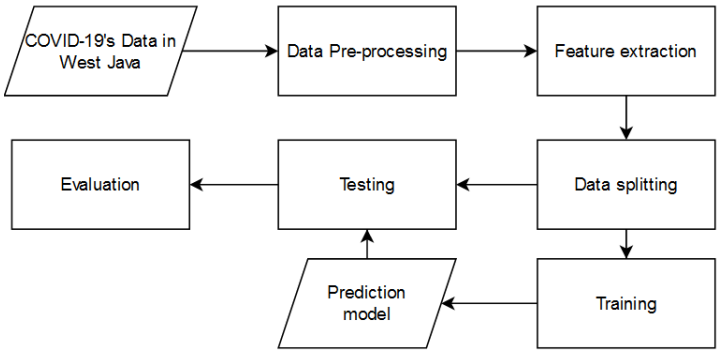


**Fig. 1.** The methodology of Time Series Prediction of Novel Coronavirus COVID-19 Data in West Java

Fig. 1 depicts the predicted number of new cases in this investigation. In the preprocessing phase, the timestamp as the date and caseNew as the number of daily new cases were incorporated. These two characteristics were retrieved via remapping, lag selection, or a combination of the two techniques. This procedure generated the following 21 categories of attributes:

a) caseNew
b) DayOfWeek
c) Weekend
d) Timestamp-remapped
e) Lag_caseNew-1
f) Lag_caseNew-2
g) Lag_caseNew-3
h) Lag_caseNew-4
i) Lag_caseNew-5
j) Lag_caseNew-6
k) Lag_caseNew-7
l) Timestamp-remapped^2
m) Timestamp-remapped ^3
n) Timestamp-remapped *Lag_caseNew-1
o) Timestamp-remapped *Lag_caseNew-2
p) Timestamp-remapped *Lag_caseNew-3
q) Timestamp-remapped *Lag_caseNew-4
r) Timestamp-remapped *Lag_caseNew-5
s) Timestamp-remapped *Lag_caseNew-6
t) Timestamp-remapped *Lag_caseNew-7

To test the system, 80 percent of the data was used as training data and the remaining 20 percent as test data. Both approaches were used to construct a model utilizing training data. The size of the second batch was 100. The kernel and noise levels of Gaussian Processes were examined in this work. In addition, the investigation focused on the size of the LMSLR random sample. The model predicted the test data to provide predictive data. The accuracy of this forecast was determined using mean absolute error (MAE) and root mean squared error (RMSE). Eq. (15) and Eq. (16) display the computations for both variables.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f(x)_i - x_i\right| \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x)_i - x_i\right)^2} \tag{16}$$

where *i, n* and $x_i$ represent index data based on timestamp order, the number of data and the i-th data, repspectively. Also $f(x_i)$ is the i-th data prediction.

## 4. Results and discussions

Multiple experiments were conducted to determine which of the two proposed procedures produced the best results. Based on MAE and RMSE, there are three parameters of the two evaluated methodologies. In addition, the optimal circumstances of both are compared. The next part describes this analysis.

### *4.1. The kernel in Gaussian Processes*

The Pearson VII function-based universal kernel, Normalized Poly Kernel, Poly Kernel, and RBF Kernel were among the kernel types evaluated. This test utilized a noise level (L) of 1.0 against a seed. The Pearson VII function-based universal kernel (PUK) yielded the minimum error compared to the other three approaches examined, as shown in Table 1. This study also revealed that the RBF kernel generated the most error. The PUK is a dependable kernel for diverse data sets with fluctuating data patterns, such as the COVID-19 data set. This approach excels at mapping data and managing diverse graph types (Zhao et al., 2016).

**Table 1**
The kernel in gaussian processes

| Kernel Type | MAE | RMSE | Kernel Type | MAE | RMSE |
|---|---|---|---|---|---|
| The Pearson VII function-based universal kernel. | 34.16 | 77.99 | Poly Kernel | 36.41 | 83.66 |
| Normalized Poly Kernel | 36.36 | 82.79 | RBF Kernel | 38.09 | 86.34 |

### *4.2. Level of Noise in Gaussian Processes*

In section 4.1, it is determined that the PUK kernel had the maximum performance. This kernel was then compared to the Gaussian noise level. Level sizes varied between 0.1 and 0.5. The inaccuracy of the PUK for various noise level measurements is displayed in Table 2. If level was equal to 0.1, the optimal situation was reached. The greater the level size, the greater the inaccuracy introduced by Gaussian processes.

**Table 2**
Level of noise in gaussian Processes

| Level of Gaussian Noise | MAE | RMSE |
|---|---|---|
| 0.1 | 23.12 | 53.43 |
| 0.2 | 26.06 | 61.13 |
| 0.3 | 27.97 | 66.09 |
| 0.4 | 29.46 | 69.35 |
| 0.5 | 30.73 | 71.73 |

**Table 3**
Sample Size in LMSLR

| Sample size | MAE | RMSE |
|---|---|---|
| 2 | 36.32 | 88.97 |
| 3 | 34.47 | 86.07 |
| 4 | 35.04 | 87.32 |
| 5 | 36.44 | 89.40 |
| 6 | 34.89 | 86.74 |
| 7 | 35.04 | 87.40 |
| 8 | 36.40 | 89.40 |
| 9 | 35.63 | 88.05 |
| 10 | 36.28 | 88.85 |

### *4.3. Sample Size in LMSLR*

Section 4.3 demonstrates the performance of the LMSLR for various sample sizes according to Table 3. The optimal results were obtained with a measure of 3 The minimum MAE was 34.47, while the maximum RMSE was 86.07. Increasing the size did not necessarily result in a reduction in error. Six samples had a smaller margin of error than five and seven samples.

## 5. Conclusion

According to Tables 1, 2, and 3, Gaussian processes performed better than LMSLR. When the Pearson VII function-based universal kernel (PUK) was the kernel and the amount of Gaussian noise was equal to 0.1, the most ideal conditions for Gaussian processes were attained. Even at the various noise levels evaluated in this study, their performance was superior to LMSLR. It demonstrated that Gaussian processes using The Pearson VII function-based universal kernel (PUK) were more resistant to flexibility in time series data under graphical settings. There were peaks in this data for COVID-19 in West Java at specific dates. Gaussian processes were best equipped to handle this circumstance. Table 4 displays the projected number of new cases for the following two weeks based on the optimal Gaussian process settings. This forecast begins on September 6, 2020, as the most recent data in this analysis is from September 5, 2020. According to Table 4, COVID-19 patients are still present in West Java, with numbers exceeding 100. It means that PSBB relaxing is theoretically not advised for West Java. This forecast, however, cannot serve as the primary reference. Because several factors have a significant impact on the number of new COVID-19 cases in Indonesia. It is insufficient to rely just on data derived from time series.

**Table 4**
Sample Size in LMSLR

| Date | Predicted number of new case | Date | Predicted number of new case |
|---|---|---|---|
| 09-06-2020* | 105 | 09-13-2020* | 105 |
| 09-07-2020* | 143 | 09-14-2020* | 143 |
| 09-08-2020* | 195 | 09-15-2020* | 190 |
| 09-09-2020* | 185 | 09-16-2020* | 177 |
| 09-10-2020* | 208 | 09-17-2020* | 191 |
| 09-11-2020* | 359 | 09-18-2020* | 324 |
| 09-12-2020* | 222 | 09-19-2020* | 196 |

## Acknowledgment

We appreciate Universitas Padjadjaran's rector. Contract No. 1735/UN6.3.1/LT/2020 provided funding for the online data and library research grant at Universitas Padjadjaran in 2020.

## References

Cheng, Y., Parker, S. T., Ran, B., & Noyce, D. A. (2016). Work Zone Crash Cost Prediction with a Least Median Squares Linear Regression Model. *Transportation Research Record, 2555*(1), 38-45.

Fadhlullah, M. U., Resahya, A., Nugraha, D. F., & Yulita, I. N. (2018, May). Sleep stages identification in patients with sleep disorder using k-means clustering. *In Journal of Physics: Conference Series, 1013*(1), p. 012162). IOP Publishing.
Gitt, A. K., Karcher, A. K., Zahn, R., & Zeymer, U. (2020). Collateral damage of COVID-19-lockdown in Germany: decline of NSTE-ACS admissions. *Clinical Research in Cardiology*, *109*(12), 1585-1587.
Guo, J., Feng, X. L., Wang, X. H., & van IJzendoorn, M. H. (2020). Coping with COVID-19: Exposure to COVID-19 and Negative Impact on Livelihood Predict Elevated Mental Health Problems in Chinese Adults. *International Journal of Environmental Research and Public Health, 17*(11), 3857.
Hayes, A. F., & Montoya, A. K. (2017). A tutorial on testing, visualizing, and probing an interaction involving a multicategorical variable in linear regression analysis. *Communication Methods and Measures, 11*(1), 1-30.
Hui, D. S., Azhar, E. I., Memish, Z. A., & Zumla, A. (2020). Human Coronavirus Infections—Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS), and SARS-CoV-2. Reference Module in Biomedical Sciences.
Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *International journal of antimicrobial agents,* 105924.
Liu, H., Ong, Y. S., Shen, X., & Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, *31*(11), 4405-4423.
Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., ... & Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery (London, England), 78*, 185.
Olivia, S., Gibson, J., & Nasrudin, R. A. (2020). Indonesia in the Time of Covid-19. *Bulletin of Indonesian Economic Studies, 56*(2), 143-174.
Pati, K. D. (2020, April). Using standard error to find the best robust regression in presence of multicollinearity and outliers. *In 2020 International Conference on Computer Science and Software Engineering (CSASE) (pp. 266-271). IEEE.*
Richardson, R. R., Osborne, M. A., & Howey, D. A. (2017). Gaussian process regression for forecasting battery state of health. *Journal of Power Sources, 357*, 209-219.
Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology, 85*, 1-16.
Sheng, H., Xiao, J., Cheng, Y., Ni, Q., & Wang, S. (2017). Short-term solar power forecasting based on weighted Gaussian process regression. *IEEE Transactions on Industrial Electronics, 65*(1), 300-308.
Toharudin, T., Caraka, R. E., Chen, R. C., Nugroho, N. T., Tai, S. K., Sueb, M., ... & Pardamean, B. (2020). *Bayesian Poisson Model for COVID-19 in West Java Indonesia. Sylwan, 164*(6), 279-290.
Weigend, A. S. (2018). *Time series prediction: forecasting the future and understanding the past*. Routledge.
Wildani, I. M., & Yulita, I. N. (2019, March). Classifying botnet attack on internet of things device using random forest. In IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002). IOP Publishing.
Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation, 46*(8), 6261-6282.
Yulita, I. N., & Wasito, I. (2013, March). gCLUPS: Graph clustering based on pairwise similarity. *In 2013 International Conference of Information and Communication Technology (ICoICT) (pp. 77-81). IEEE*.
Yulita, I. N., Fanany, M. I., & Arymurthy, A. M. (2017, September). Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification. *In 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 1-6). IEEE.
Yulita, I. N., Fanany, M. I., & Arymurthy, A. M. (2017, October). Sleep stage classification using convolutional neural networks and bidirectional long short-term memory. *In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 303-308). IEEE*.
Yulita, I. N., Hidayat, A., Abdullah, A. S., & Awangga, R. M. (2018). Feature extraction analysis for hidden Markov models in Sundanese speech recognition. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 16*(5), 2191-2198.
Yulita, I. N., Julviar, R. R., Triwahyuni, A., & Widiastuti, T. (2019, July). Multichannel Electroencephalography-based Emotion Recognition Using Machine Learning. *In Journal of Physics: Conference Series* (Vol. 1230, No. 1, p. 012008). IOP Publishing.
Zhao, J., & Han, M. (2016). An efficient model for the prediction of polymerisation efficiency of nano-composite film using Gaussian processes and Pearson VII universal kernel. *International Journal of Materials and Product Technology, 52*(3-4), 226-237.
Zuhairoh, F., & Rosadi, D. (2020). Indonesian Journal of Science & Technology. *Indonesian Journal of Science & Technology, 5*(3), 456-462.